Summary Report on Lead Score Case Study

To address the problem of selecting the most promising leads for X Education Company, we built a Logistic Regression model on the given dataset. Our approach consisted of the following steps:

1) Data loading and Cleaning:

We loaded the data into a Jupyter notebook and performed missing value and outlier treatment on the dataset. We also dropped columns that had more than 45% missing values. The columns having single unique entries dominantly were also discarded.

2) Exploratory Data Analysis (EDA):

Analysis of each column were carried out with Univariate analysis. Further, Bivariate analysis with the target variables were also carried out to understand the conversion rate of each feature. For numerical variables multivariate analysis with heat map was also carried out. This gave us a fair idea about the impact of different variables on the target variable.

3) Data Preparation and Modelling:

- *Dummy Variable Creation:* Dummy variables for all the categorical columns except the binary one (yes/no) were created.
- *Train-test split and feature scaling:* We performed train-test split (70:30) using the sklearn library and scaled the numerical variables of train data using Standard Scaler.
- *Feature selection:* We used Recursive Feature Elimination (RFE) to automatically select 20 features for initial model.
- Logistic Regression: We used the statsmodels GLM method to perform Logistic Regression on the selected features and checked coefficients, p-values, and Variance Inflation Factor (VIF).
- Model evaluation: We evaluated the model using accuracy, specificity, and sensitivity. We also plotted the Receiver Operating Characteristic (ROC) curve to check the balance between True Positive Rate (TPR) and False Positive Rate (FPR).
- Cut-off selection: We selected the optimal cut-off point (0.34) for lead conversion by plotting Accuracy, Sensitivity and Specificity for different probability cut-offs.
- Re-evaluation: We re-evaluated the model using the selected cut-off point and found an accuracy of 90.6%, precision of 86.4%, and recall of 90%.
- Model testing: We scaled the test data and performed prediction of lead conversion. We evaluated the prediction on test data by accuracy, specificity and sensitivity matrix, and we found accuracy 91.1%, precision 85.6% and recall 91.9%.
- Lead Score Calculation: We created lead conversion score = (conversion probability * 100) to give a score between 0 to 100 where higher the value means the lead is “hot” and there is high possibility that the lead can be converted.

Observation:

- Top impactful features identified in our model are:
  - Tags_Closed by Horizzon
  - Tags_Already a student
  - Tags_Ringing
  - Tags_invalid number
  - Lead_Source_Welingak Website

- Apart from these EDA also provides some insights about the conversion rates. Conversion rate is very high for Lead Add Form as Lead Origin. For Reference and Welingak website the conversion rate is very high. For working professionals’ conversion rate is maximum. Converted leads have higher median Total time spent on the website than non-converted one. Company X can surely look into all these factors for targeting appropriate leads.